Ensemble Machine Learning Approach for Diabetes Prediction

1K.R. SriPreethaa, 2N. Yuvaraj and 3G. Jenifa
1,2,3Department of Artificial Intelligence and Data Science
1,2,3KPR Institute of Engineering and Technology Coimbatore, India.
1k.r.sripreethaa@kpriet.ac.in, 2yuvaraj.ni@kpriet.ac.in, 3jenifa.gj@kpriet.ac.in

Abstract - The technological advancements applied in the area of healthcare systems helps to meet the requirement of increasing global population. Due to the infections by the various microorganisms, people around the world are affected with different types of life-threatening diseases. Among the different types of commonly existing diseases, diabetes remains the deadliest disease. Diabetes is a major cause for the change in all physical metabolism, heart attacks, kidney failure, blindness, etc. Computational advancements help to create health care monitoring systems for identifying different deadliest diseases and its symptoms. Advancements in the machine learning algorithms are applied in various applications of the health care systems which automates the working model of health care equipment’s and enhances the accuracy of disease prediction. This work proposes the ensemble machine learning based boosting approaches for developing an intelligent system for diabetes prediction. The data collected from Pima Indians Diabetes (PID) database by national institute of diabetes from 75664 patients is used for model building. The results show that the histogram gradient boosting algorithms manages to produce better performance with minimum root mean square error of 4.35 and maximum r squared error of 89%. Proposed model can be integrated with the handheld biomedical equipment’s for earlier prediction of diabetes.

Keywords - diabetes, healthcare systems, ensemble machine learning algorithms, prediction, predictive analysis.

1. Introduction
Health care systems have become vital agenda globally. Health care system are designed for need of the people. The growing population demands the accurate exacerbation of chronic illness [1]. The health problems requires quick remedial actions. Health services aims at providing improved quality health, early diagnosis, treatment, rehabilitation of the diseased people. The various objectives of health care services are effective interventions, prevention, cure and rehabilitation of individuals or population [2]. In modern health care systems the valuable information plays a critical role in providing efficient health system [3]. Information and communication technologies had developed health care systems to get improved through standardization of health information. Computer based diagnosis systems, high level monitoring and also educating the society on health related awareness [4]. The automated hospital information system can provide enhanced the quality of care and treatment even at remote areas.

Among various other chronic disease, majority of population is found affected by diabetes. Diabetes is the inability of the pancreas to produce insulin or effectively utilize the insulin produced by it. Insulin is the hormone that regulates the level of sugar in blood [5]. Hyperglycemia is the raised level of blood sugar causing diabetes that affect the vital organs of the human body system especially nerves and blood vessels. A survey report in 2014, claims that 8.5% of adults with age 18 were affected by diabetes [6]. Additional, in 2015 diabetes was reported to cause 1.6 million deaths directly. In United states it was estimated there is an increase in diabetic population from 108 million in 1980 to 422 million in 2014. The most vulnerable age of the diabetic patients was above 18 years of age [7]. The disease is found prevalently increasing in countries with an average or low income. Diabetes causes blindness, heart attacks, stroke and also kidney failures. Almost 50% of all deaths before age 70 are due to high blood sugar [8]. The World health organization (WHO) claimed that in 2030 diabetes would be the seventh death causing disease. The National Diabetes Statistics Report provides statistical updates about the diabetes in United States to the scientific community [9]. The information includes the prevalence of diabetes, pre-diabetic symptoms, risk factors, acute and long term complications, death and costs. The report in 2015 shows about 84.1 million people at the age of 18 had developed pre-diabetes. Nearly 48.3% of the adults with age 65 and above had developed diabetes [10]. It is necessary to build a system that predicts the diabetes accurately and effectively. Development in secured wireless body area network paved a way for health care systems for monitoring the disease [11]. Advancements in big data, made medical field applications to the focus on health care industries to develop an enhanced predictive diabetic diagnostic model [12].

The models integrated with Genetic algorithms (GA) and artificial neural networks (ANN) were built to predict the variable based on the relationship between variables in the training data [13]. Many predictive models were built using gene expression programming (GEP) like finding the bond strength in concrete structures using shear capacity [14]. ANNs were used for predicting the shear capacity of fibre reinforcement polymers beams without using shear capacity [15]. The Back propagation neural networks (BPNN) was implemented in predicting the shear resistance of FRP-retrofitted RC beams. A linear and non-linear based multiple regression models were investigated by group method of data handling (GMDH) networks for predicting bond strength [16]. Th5 GMDH model was proved to be accurate and out performing. An ANN based and fuzzy logic were used to find the heart attacks based on the medical dataset out [17]. A multi-gene genetic
Boosting algorithms help in improving the performance of ensemble learning models without bias. This paper aims to build a diabetic predictive model based on ensemble learning.

The ensemble algorithms have proved to enhance the performance over individual machine learning algorithms due to the integration of operating principles. Boosting algorithms help in improving the performance without bias. This paper aims to build a diabetic predictive model based on ensemble learning.

The flow of the paper is organized as follows: The section 1 provided the introduction to health systems and discussion on the work related to proposed healthcare systems. The section 2 provides the overview of the proposed method. The section 3 discusses the internal working of the algorithms used followed by the evaluation measures and results in section 4. Finally, section 5 gives the conclusion and future of the proposed prediction methodology.

2. Overview of the proposed method

The proposed work concentrates on developing an intelligent prediction model for diabetes prediction. The entire model building process involves two phases namely training the model and testing the model (Figure 1).

3. Ensemble Methods for enhanced prediction

Random forest

Random forest (RF) is most widely used ensemble learning method for regression and classification. RF is considered to be a best classifier due to its simplicity and effectiveness. RF ensemble classifier uses a bagging for ensemble method and decision tree for individual model. RF provides difference among the varied decision trees of the forest through bootstrap sampling. Each decision tree is trained with different training dataset by replacing the original dataset. During growth of each tree, RF includes the extra boost of diversity. From the total set of variables available in the data, a subset of variables is considered to select the best cut for tree branching. The selection is made based on the square root of total set of variables. The depth of the tree is allowed to grow without pruning. Based on the majority of voting among the decision trees the final prediction is made. The accuracy of the model remains the same even if large portion of the data is missing.

Bootstrap aggregation is the general training algorithm used for training the random forest. The given data set is divided into X and Y. A training set X = x1, ..., xn with responses Y = y1, ..., yn, bagging repeatedly (B times) selects a random sample with replacement of the training set and fits trees to these samples. For b = 1, ..., B: Sample, with replacement, n training examples from X, Y; call these Xb, Yb. Train a classification or regression tree fb on Xb, Yb. After training, predictions for unseen samples x can be made by averaging the predictions from all the individual regression trees on x using the Equation (1).

\[ f^* = \frac{1}{B} \sum_{b=1}^{B} f_b(x') \]  

(1)

Histogram Gradient Boosting

Gradient Boosting is an additive model based ensemble classifier where each iteration is done at the deepest descent minimization with the loss function. Numerical optimization is done in function space to estimate the predictive function. Both linear regression and decision tree are used as a base learner. Initial guess F0(x) is calculated for all output samples. Gradient boosting repeats fitting the tree till the maximum number of estimators is reached. The output x is predicted by using the initial guess value and the scaled output through ensemble. A negative binomial log function is used to build a multiclass classification model. For multiclass problems, HGB approximates the additive function Fk(x) for each class label l by loss function shown in Equation (2).

\[ L(\{y_1, F_k(s)\}_{i=1}^{n}) = -\sum_{i=1}^{n} \log y_i p_l(x) \]  

(2)

where L number of classes, y1 takes value 1 when sample x belongs to the class 0 or 1. Pl(x) is the probability of x for the class l.

Proposed model for diabetes prediction

The proposed model shown in Figure 2 for diabetes prediction takes the training data as an initial input and
random number of subset is generated for n number of training data. Trees are generated from the random subset subsequently and splitting of tree nodes is done. Ensemble classifiers namely random forest and HGBoost classifier algorithm.

The aggregated results of the trees are used for model building the diabetes prediction model. The proposed model is validated with test data to predict the diabetes possibility of the input data. The model is evaluated on proposed ensemble classifier approaches and the results shows that HGBoost manages to attain better performance with greater accuracy than the random forest.

4. Results and Discussions
The overall performance of the proposed ensemble methods was compared using different performance metrics like root mean square error (RMSE), R-square error, covariance (COV) and integral absolute error (IAE). The performance comparison of the proposed ensemble methods for diabetes prediction is represented in the Table 1.

Evaluation Metrics
The different evaluation metrics used and the formula to evaluate the same is explained in the following section. RMSE is the root mean square standard deviation of the feature values available in the entire dataset. RMSE helps to measures the magnitude of the error. RMSE is calculated as the square root of average of squared difference between predicted diabetic observation and the actual diabetic observation. RMSE is calculated using below Equation (3).

\[ RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \]  \hspace{1cm} (3)

where \( n \) is the total number of observations, \( y_i \) denotes predicted and \( \hat{y}_i \) denotes actual diabetic observation. R-squared is one of the statistical measures, which represents the proportion of the variance of a dependent variable in the feature set with the independent variable. This correlation explains the strength of the relationship between an independent and dependent variable that were used for the model building. R-squared is calculated using the formula given in Equation (4).

\[ R^2 = 1 - \frac{\sum_{i}(y_i - \hat{y}_i)^2}{\sum_{i}(y_i - \bar{y})^2} \]  \hspace{1cm} (4)

where, \( Y \) is the actual value, \( y_i \) is the predicted value of \( y \) and \( \bar{y} \) denotes the mean of \( y \) values. Covariance measure (Equation 5) is used to evaluate the total variation in the actual and predicted diabetes values. The actual and the predicted values are closer to each other when the observation is positive. When the relationship is negative, it represents the worst fit of the actual and predicted observations.

\[ \text{cov}(x, y) = \frac{\sum_{i}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \]  \hspace{1cm} (5)

Where, \( x_i \) denotes the predicted values, \( y_i \) denotes the actual values, \( \bar{x} \) and \( \bar{y} \) denotes the mean of predicted and mean of observed values out of the available \( n \) number of observations. IAE measures the model performance by integrating the absolute error over predicted and actual observations from the diabetes dataset. IAE is calculated using the formula given in Equation (6).

\[ IAE = \int_{0}^{\infty} |e(t)| dt \]  \hspace{1cm} (6)

The visualization of the performance analysis of the proposed ensemble methods is represented in Figure 3.

<table>
<thead>
<tr>
<th>Performance Measures</th>
<th>Ensemble Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Random Forest</td>
</tr>
<tr>
<td>RMSE</td>
<td>4.73</td>
</tr>
<tr>
<td>R-Square</td>
<td>0.19</td>
</tr>
<tr>
<td>COV</td>
<td>0.33</td>
</tr>
<tr>
<td>IAE</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Table 1: Performance Comparison
5. Conclusion

Healthcare system plays a vital role in monitoring the daily health aspects of the humans all around the globe. The advancements in computational algorithms render a massive support in developing intelligent healthcare support applications. The health care support systems for early prediction of the diabetes is the need of the day. In this proposed work, the advancements in the ensemble methods were applied for building diabetes prediction system. HGBoost method manages to produce the better performance over the random forest algorithm with minimum RMSE of 4.35, minimum IAE of 0.17, minimum covariance of 0.29 and maximum R-Square error of 0.89. As a future work, maximum attention can be given on improving the accuracy further and there by integrating the proposed model with the handheld devices for diabetes prediction.

References


